

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-157091

(43)Date of publication of application : 31.05.2002

(51)Int.Cl.

G06F 3/06

G06F 12/16

(21)Application number : 2000-353010

(71)Applicant : HITACHI LTD

(22)Date of filing : 20.11.2000

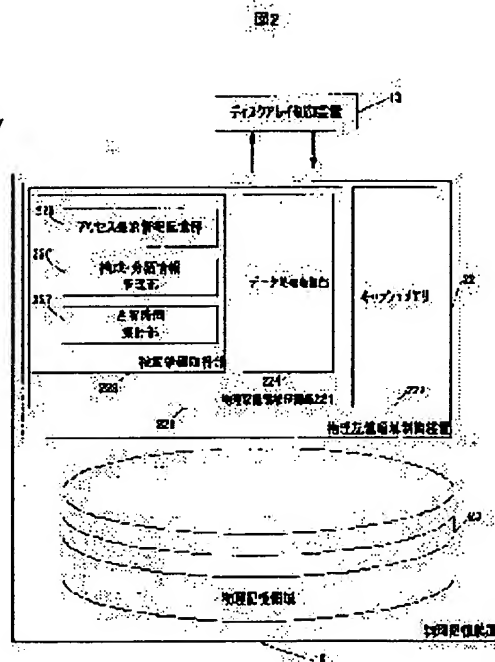
(72)Inventor : EGUCHI KENTETSU
MOGI KAZUHIKO
ARAKAWA TAKASHI
OEDA TAKASHI
ARAI HIROHARU

(54) STORAGE SUB-SYSTEM, AND MEMORY USED THEREFOR

(57)Abstract:

PROBLEM TO BE SOLVED: To obtain an occupied time of a logic storage area in a physical memory, and to obtain precise access occupied time information in every I/O to the physical memory.

SOLUTION: A physical storage area controller 22 on the individual physical memory 15 is provided with a table 225 for storing information about access requirement from a host computer, a table 227 for totalizing the occupied time as to access, a table 226 for control information for classifying constitution of a disk array, and a data processing control part 224 for obtaining constitution information and classification information of the logic storage area from a disk array controller 13, and for requesting the constitution information and the classification information of the logic storage area to the disk array controller, when necessary. The disk array controller 13 is provided with a means for transmitting the constitution information of the disk array at the present time to the physical storage area controller in response to the request from the physical storage area controller on the physical memory.



LEGAL STATUS

[Date of request for examination]

18.07.2003

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision
of rejection]

[Date of requesting appeal against examiner's
decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

* NOTICES *

JPO and NCIP are not responsible for any damages caused by the use of this translation.

1. This document has been translated by computer. So the translation may not reflect the original precisely.
2. *** shows the word which can not be translated.
3. In the drawings, any words are not translated.

CLAIMS

[Claim(s)]

[Claim 1] A means to be connected to 1 or two or more computers, and to acquire the operating condition information on two or more physical memory equipments and two or more of these physical memory equipments, in the storage subsystem which has a means to perform matching with the logic storage region which said computer makes a read/write object, and the physical memory field of said physical memory equipment it is the storage subsystem which each of two or more of said physical memory equipments is equipped with a physical memory field control unit, and is characterized by this physical memory field control unit having a means to acquire the operating condition of a physical memory field.

[Claim 2] A means to be connected to 1 or two or more computers, and to acquire the operating condition information on two or more physical memory equipments and two or more of these physical memory equipments, in the storage subsystem which has a means to perform matching with the logic storage region which said computer makes a read/write object, and the physical memory field of said physical memory equipment A means to acquire the operating condition information on said two or more physical memory equipments, and a means to perform matching with the logic storage region which said computer makes a read/write object, and the physical memory field of said physical memory equipment it is prepared in the control unit which controls two or more above-mentioned physical memory equipments. Said control unit Furthermore, it has a means to transmit the information which performed matching with the logic storage region of physical memory equipment, and the physical memory field of physical memory equipment to said two or more physical memory equipments of each. It is the storage subsystem which each of two or more of said physical memory equipments is equipped with a physical memory field control unit, and is characterized by this physical memory field control unit having a means to acquire the operating condition of a physical memory field.

[Claim 3] It is physical memory equipment which is equipped with a physical memory field control unit in the physical memory equipment which constitutes a storage subsystem, and is used for the storage subsystem according to claim 1 or 2 characterized by this physical memory field control unit having a means to acquire the operating condition of a physical memory field.

[Claim 4] Said physical memory field control unit is physical memory equipment according to claim 3 characterized by having further a means to store the operating condition information on the acquired physical memory field.

[Claim 5] Said physical-memory field control unit is physical-memory equipment according to claim 3 or 4 characterized by to have further a means store the information which matched the logic storage region and the physical memory field of the physical-memory equipment which receives the operating condition information on the physical memory field of self-physical memory equipment from a means to transmit to said control unit, and said control unit, according to the acquisition demand of the operating-condition information on a physical-memory field which receives from said control unit.

[Translation done.]

* NOTICES *

JP0 and K0P1 are not responsible for any damages caused by the use of this translation.

- 1.This document has been translated by computer. So the translation may not reflect the original precisely.
- 2.*** shows the word which can not be translated.
- 3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]

[0001] [Field of the Invention] This invention relates to the store used for a storage subsystem and its system, and relates to the store used for the storage subsystem which has two or more stores especially, and its system.

[0002] [Description of the Prior Art] The disk array system is known as a highly efficient secondary-storage system used for a computer system.

[0003] A disk array system is a system which made it possible to perform read/write of the data which arrange two or more physical memory equipments in the shape of an array, divide and store data in each physical memory equipment, and said each physical memory equipment is operated to juxtaposition, and are stored in said each physical memory equipment by dividing at a high speed.

[0004] As a conventional technique about a disk array system, the technique indicated by D.A.Patterson, G.Gibson, and R.H.Kats, "A Case for Redundant Arrays of Inexpensive Disks (disk array) etc." (inch Proc-ACM SIGMOD, pp.109-116, and June 1988), etc. is known. This conventional technique gives the classification of level 5 from level 1 to the disk array system which added redundancy according to that configuration. Moreover, a disk array system without redundancy may be added to these classification, and this may be called level 0 to them. In order to realize each above-mentioned level as a different configuration according to redundancy etc., cost differs from performance characteristics etc. And in building a disk array system, the array (group of physical memory equipment) of two or more level is made intermingled in many cases. Here, the group of the disk array which added redundancy is called a parity group. Moreover, in order to realize optimal cost performance in cost's changing with the engine performance, capacity, etc. also about physical memory equipment, and building a disk array system, two or more sorts of physical memory equipments with which the engine performance differs from capacity too may be used.

[0005] It distributes to said above physical memory equipments, and the data stored in a disk array system are arranged. For this reason, a disk array system needs to perform matching of the physical memory field which shows the logic storage region which the host computer connected to a disk array system accesses, and the storage region of said physical memory equipment, i.e., address translation.

[0006] The technique indicated by JP,9-274544,A etc. is known as a conventional technique about the disk array system which processes address translation. This conventional technique says that a means to acquire the information about I/O access over the logic storage region from a host computer, and a means to change matching with the physical memory field of a logic storage region, and to perform physical relocation realize the optimal arrangement of the stored data, and the technique in which a disk array control unit acquires the I/O access occupancy hour entry to a logic storage region is indicated by this official report.

[0007] In order to perform the load distribution of a disk array system, in case it changes matching with the physical memory field of a logic storage region and performs physical

relocation, since the I/O access occupancy hour entry to a logic storage region serves as data which become origin, it is important.

[0008]

[Problem(s) to be Solved by the Invention] Although, as for the conventional technique indicated by the official report mentioned above, it is said that a disk array control unit acquires the occupancy time amount by I/O to a logic storage region, the approach shown in this conventional technique has the trouble that it explains below.

[0009] First, the case where the writing (light) of data is performed to a certain logic storage region is considered. In this case, the light of data is performed to the physical memory field corresponding to the logic storage region concerned. A physical memory field is in physical memory equipment, and physical memory equipment is constituted by the cache memory and the physical memory field which mainly carry out the cache of a physical memory field control section and the data. And when carrying out the light of the data to a physical memory field, it writes in, when a physical memory field control section writes light data in a cache, and the response of termination is notified to a disk array control unit. For this reason, in order that the conventional technique mentioned above may actually carry out the light of the data, it will have the trouble that the time amount which accessed the physical memory field is not known.

[0010] Next, the case where reading (lead) of data is performed to a certain logic storage region is considered. In this case, in fact, although the lead of the data in the physical memory field corresponding to the logic storage region concerned is performed, when that lead data is in the cache memory in physical memory equipment, a physical memory field is not accessed, but cache memory is accessed, and that data is returned. For this reason, it will have the trouble that the exact time amount which could not distinguish and accessed the physical memory field does not understand whether the conventional technique mentioned above actually had access in the physical memory field for the data lead.

[0011] Moreover, the case where the logic storage region currently crossed to A, B, C, D, and multiple times has access is considered. And the response of A' and B is made into B' etc. for the response of A, and time of day of B (t) and response A' is made [the time of day of Access A] into A' (t) etc. for the time of day of A (t) and Access B. Here, Access A accesses physical memory equipment with a data lead, leads data, and assumes Access B to be what had the response B' without accessing physical memory equipment, since data are in a cache with a data lead. In this case, the direction of response B' of Access B which came after response A' of the access A which suited previously becomes previously. That is, it becomes A(t) < B (t), A' (t) > B' (t). At this time, the conventional technique which acquires the occupancy time amount by I/O to a logic storage region in a disk array control unit produces the trouble that sweet red bean soup with mochi cannot perform whether which I/O accessed the physical memory field in physical memory equipment how much, or it hit into the cache of physical memory equipment.

[0012] The purpose of this invention is by solving the trouble of the conventional technique mentioned above and acquiring the occupancy time amount of a logic storage region with physical memory equipment to offer the store which offers the storage subsystem which enabled it to acquire the occupancy time amount (real operating time) of each logic storage region by the system configuration unacquirable [with a disk array control unit], and is used for this.

[0013] The purpose of this invention moreover, by taking into consideration the effect on the physical memory equipment for every I/O in order to be able to make small the error of the analysis of the utilization factor of the logic storage region of said disk array system, or utilization factor prediction and to perform optimum performance tuning more it is in offering the store which offers the storage subsystem which can acquire the access occupancy hour entry for every I/O to physical memory equipment with a more high precision, and is used for this.

[0014]

[Means for Solving the Problem] According to this invention, said purpose is connected to 1 or two or more computers. Two or more physical memory equipments, in the storage subsystem which has a means to acquire the operating condition information on two or more of these physical memory equipments, and a means to perform matching with the logic storage region which said computer makes a read/write object, and the physical memory field of said physical

memory equipment A means to acquire the operating condition information on said two or more physical memory equipments, and a means to perform matching with the logic storage region which said computer makes a read/write object, and the physical memory field of said physical memory equipment. It is prepared in the control unit which controls two or more above-mentioned physical memory equipments. Said control unit. Furthermore, it has a means to transmit the information which performed matching with the logic storage region of physical memory equipment, and the physical memory field of physical memory equipment to said two or more physical memory equipments of each, it has a physical memory field control unit, and this physical memory field control unit is attained by having a means of two or more of said physical memory equipments to acquire the operating condition of a physical memory field, respectively. [0015] Moreover, said purpose is set to the physical memory equipment which constitutes a storage subsystem. A means by which have a physical memory field control unit and this physical memory field control unit acquires the operating condition of a physical memory field. An acquisition demand of the operating condition information on a means to store the operating condition information on the acquired physical memory field, and the physical memory field received from said control unit is accepted. It is attained by having a means to transmit the operating condition information on the physical memory field of self-physical memory equipment to said control unit, and a means to store the information which matched the logic storage region and physical memory field of the physical memory equipment received from said control unit. [0016]

[Embodiment of the Invention] Hereafter, a drawing explains the operation gestalt of the storage used for the storage subsystem by this invention, and its system to a detail.

[0017] The block diagram showing the configuration of the computer system with which drawing 1 was equipped with the storage subsystem by this invention, and drawing 2 are the block diagrams showing the configuration of physical memory equipment. In drawing 1 R> 1 and drawing 2 a host and 12 10 A storage subsystem, 13 disk array control information and 15 for a disk array control device and 14 Physical memory equipment, in 16, a disk array and 17 an I/O bus and 19 for a control terminal and 18 A network, 22 a physical memory field and 130 for a physical memory field control unit and 23 The read/write processing section, 131 the relocation decision processing section and 133 for the operating condition information acquisition processing section and 143 The relocation executive operation section, 141 class configuration information and 143 for the information corresponding to logic/physics, and 142 Class attribute information, 144 physical field operating condition information and 146 for logic field operating condition information and 145 Relocation decision horizon information, 147 free-space information and 149 for relocation activation time information and 148 Relocation information, 14A — a storage occupancy hour entry and 221 — for the operation information acquisition section and 224, as for the access request information storage section and 226, a data-processing control section and 225 are [a physical memory field control section and 222 / cache memory and 223 / configuration / classification Research and Data Processing Department and 227] the occupancy time amount total sections.

[0018] The computer system shown in drawing 1 consists of 1 which is the calculating machine of a high order or two or more hosts 10, a storage subsystem 12, and a control terminal 17. It connects with the storage subsystem 12 by I/O bus 18, and a host 10 publishes lead of data, and I/O for light processing to the storage subsystem 12. In case this I/O is performed, a host 10 specifies the logical storage region of the storage subsystem 12. That is, a host 10 accesses with the address of a usually logical storage region to the data in a storage subsystem. Moreover, I/O bus 18 is constituted by ESCON, SCSI, the fiber channel, etc.

[0019] The storage subsystem 12 consists of a disk array control unit 13 and two or more physical memory equipments 15. The disk array control unit 13 is equipped with the read/write processing section 130, the operating condition information acquisition processing section 131, the relocation decision processing section 132, and the relocation executive operation section 133, and these processing sections process read/write processing, operating condition information acquisition processing, relocation decision processing, relocation executive operation, etc. Moreover, the disk array control non-dense of the storage subsystem 12 holds the

information 141 corresponding to a logic storage region / physical memory field, the class configuration information 142, the disk array configuration information 1400 of class attribute information 143 grade, storage occupancy hour entry 14A of the logic field operating condition information 144 and physical field operating condition information 145 grade, the relocation decision horizon information 146, the relocation activation time information 147, the free-space information 148, and relocation information 149 grade. In addition, others and parity group information, RAID level information, etc. which were mentioned above may be included in the disk array configuration information 14 mentioned above. [information] [0020] Moreover, the host 10, the disk array control unit 13, and the control terminal 17 are mutually connected by the network 19. A network 19 may be constituted by Ethernet (trademark), FDDI, the fiber channel, etc. A control terminal 17 is usually used in order to perform maintenance, management, etc. of the storage subsystem 12.

[0021] Moreover, in explanation of the operation gestalt of this invention, although it exists, respectively, since it is not important, the component which surely exists in computers, such as memory for performing processing which appears in a host 10, the disk array control unit 13, and a control terminal 17, respectively, and CPU, is not specified here.

[0022] The class division of two or more physical memory equipments 15 formed in the above-mentioned storage subsystem 12 is carried out for every engine performance of physical memory equipment, and they constitute the disk array 16 for every class. Moreover, although not shown clearly, two or more physical memory equipments are used, and the parity group is constituted here. And each of physical memory equipment 15 is constituted by the physical memory field control device 22 which controls the physical memory field 23 and this physical memory field 23 to be shown in drawing 2, and various data are stored in the physical memory field 23.

[0023] Moreover, if the address of the physical memory field 23 seems to have mentioned above directly from a host 10, it does not break, but a host 10 accesses the data on two or more logical storage regions on two or more physical memory fields 23. Namely, a host 10 accesses by specifying a logic storage region to the data in the storage region of each physical memory equipment 15 in the storage subsystem 12.

[0024] The disk array control unit 13 is connected with two or more physical memory equipments 15. The lead and light processing instruction I/O which controlled two or more physical memory equipments 15, or were emitted by said host 10 Make the address of the logic storage region where the appointed data exist, and the address of a physical memory field with the address of the logic storage region correspond, transmit data I/O to suitable physical memory equipment 15, and if it is light processing The data transmitted by the host 10 are transmitted to physical memory equipment 15, and if it is lead processing, the data transmitted from physical memory equipment 15 are received, and it is processing transmitting to a host 10 etc.

[0025] The physical memory field control unit 22 which it has in physical memory equipment 15 is constituted by the physical memory field control section 221 and cache memory 222. Cache memory 222 has the quick rate of processing of the read/write of data compared with the physical memory field 23. And cache memory 222 is used as follows about the data about the lead or light instruction transmitted from the disk array control device 13. That is, in case the light data transmitted from the disk array control device 13 are written in the physical memory field 23 in light processing, data are written also in cache memory 222. Moreover, in case data reading appearance is carried out from the physical memory field 23 in lead processing, when it was written in cache memory 222, or the same data are in cache memory and it comes to physical memory equipment from the disk array control device 13 as a lead instruction to the data by former lead processing, the read data do not read the data from the physical memory field 23, but read it from cache memory 222. Thereby, the processing engine performance of physical memory equipment 15 can be improved.

[0026] The physical memory field control section 221 is mainly equipped with the operation information acquisition section 223 and the data-processing control section 224, and is constituted. The data-processing control section 224 receives the lead or light instruction of data transmitted from the disk array control device 13. And the data-processing control section

224 will read the lead data from cache memory 222, if cache memory 222 is accessed and the lead data exists in cache memory 222, when the received instruction is a lead instruction, if the data is not in cache memory 222, it will access the physical memory field 23, will read the lead data, and will transmit data to the disk array control device 13. Moreover, the data-processing control section 224 is after that, and writes the data in the physical memory field 23 at the same time it writes the data transmitted from the disk array control device 13 in cache memory 222, when the received instruction is a light processing instruction. You may not write light data in cache memory 222, but may also write them in the direct physical memory field 23.

[0027] The operation information acquisition section 223 is constituted by the access request information storage section 225, configuration / classification Research and Data Processing Department 226, and occupancy time amount total section 227 grade. When a logic storage region with the data specified by the I/O process, the physical memory field 23 where the data exists, or cache memory 222 is accessed, the above-mentioned data-processing control section 224 classifies the hour entry of the access for every (random access, sequential access, etc.) processing classification of an I/O process, and records it on the occupancy time amount total section 227. Moreover, from the disk array control device 13, the data-processing control section 224 receives information, such as correspondence information on the address of the logic storage region in a disk array, and the address of the physical memory field 23, and engine performance of physical memory equipment 15, and records it on configuration / classification Research and Data Processing Department 226. Furthermore, the data-processing control section 224 receives a lead of data or light instruction data transmitted from disk array control device 13 grade, and records the instruction data on the access request information storage section 225 so that it may be possible to receive two or more I/O processes.

[0028] The logic storage region according to an I/O process within physical memory equipment 15 when physical memory equipment 15 has a configuration which was mentioned above, the physical memory field 23. Or it becomes possible to classify the hour entry at the time of access when accessing cache memory 222 for every (random access, sequential access, etc.) processing classification of an I/O process, and to record on the occupancy time amount total section 227. It classifies into a physical memory field to what whether time amount access was carried out and physical memory equipment 15 of cache memory 22 it hit according to an I/O process, and it becomes possible to total the occupancy time amount.

[0029] Drawing 3 is a flow chart, explaining processing actuation of a disk array control device when a storage subsystem is started, and explains this hereafter.

[0030] (1) The disk array control unit 13 transmits the information 141 corresponding to the logic/physics which is the matching information on the address of the logic storage region in the physical memory field 23 in physical memory equipment 15, and the address of the physical memory field where the logic storage region actually exists, the class configuration information 142, and the disk array configuration information 14 of class attribute information 143 grade to the physical memory equipment 15 connected with self-equipment 13 at the time of starting of the storage subsystem 12 (steps 300 and 310).

[0031] (2) Next, the disk array control unit 13 receives the notice the physical memory equipment 15 sent from the physical memory field control unit 22 changed [the notice] to the accessible ready state, when physical memory equipment 15 becomes accessible by transmission of the information mentioned above. Things come and physical memory equipment 15 is in the condition of the initialization termination by the disk array configuration information 14 (step 320).

[0032] (3) Then, the disk array control device 13 receives data with more various what has transmitted host I/O of a lead or light processing to the storage subsystem 12 to the logic storage region in the storage subsystem 12, things which deliver an instruction and data with disk array control devices than a host 10 by I/O-bus 18 course (step 330).

[0033] (4) When host I/O is received as said received data, the disk array control unit 13 receives the lead or light demand to the logic storage region specified by host I/O, and asks for the logic storage region address and the corresponding address of the physical memory field 23 using the information 141 corresponding to the logic/physics which changes the address (logical

address) of the logic storage region into the address (physical address) of a physical memory field (steps 340 and 350).

[0034] (5) The disk array control device 13 specifies the address of the physical-memory field where predetermined data exist, and, in lead processing, transmits light data to the physical-memory equipment which transmits lead data for the physical memory equipment which has the above-mentioned physical address to lead data to read-out and a host 10, receives the light data transmitted by the host 10 in light processing, and has the physical address (360).

[0035] Drawing 4 is a flow chart, explaining processing actuation of a disk array control device when correspondence of the address of a logic storage region and the address of a physical memory field changes, and explains this hereafter.

[0036] (1) By the change in physical memory equipment 15, change of RAID level, and a logic storage region moving the disk array control device 13 to the address of a physical memory field different from a certain physical memory **** address now etc. When it supervises that correspondence of the address of a logic storage region and the address of a physical memory field changed and is changeable, physical memory equipment 15 is received again. The information 141 corresponding to the logic/physics which is the matching information on the address of the logic storage region in the physical memory field 23 in physical memory equipment 15, and the address of the physical memory field where the logic storage region actually exists, The class configuration information 142 and the disk array configuration information 14 of class attribute information 143 grade are transmitted (step 3101).

[0037] (2) Next, the disk array control unit 13 receives the notice the physical memory equipment 15 sent from the physical memory field control unit 22 changed [the notice] to the accessible ready state, when physical memory equipment 15 becomes accessible by transmission of the information mentioned above. Things come and physical memory equipment 15 is in the condition of the updating termination by the disk array configuration information 14 (step 3201).

[0038] (3) Subsequent processing is performed like the case of steps 3300, 3400, 3500, and 3600 explained by drawing 3 (steps 3301, 3401, 3501, and 3601).

[0039] In addition, in steps 320 and 3201 mentioned above, as for the disk array control unit 13, ** does not need to receive the information that the physical memory field control unit 22 to physical memory equipment 15 changed in the accessible condition. In this case, what is necessary is just to perform access processing for performing predetermined processing to physical memory equipment 15, when it assumes that it is in accessible conditions, such as a lead and a light, to physical memory equipment 15 and the access directions to a certain storage region come for the disk array control unit 13 after a certain fixed time amount. Moreover, when there is no response into waiting and fixed time amount until it performs access processing in order to perform predetermined processing again when there is no response to predetermined access processing, or a response comes back, it is good also as a method which tells that to the module which issued the access directions to a certain storage region.

[0040] Moreover, the information 141 corresponding to logic/physics in the above-mentioned information to which a logic storage region and a physical memory field are made to correspond. And the logical address is the address which shows the logic storage region which a host 10 uses in said read/write processing section 130. Moreover, a physical address is the address which shows the field on the physical memory equipment 15 with which data are actually stored, and consists of the physical memory device number and the address in physical memory equipment. A storage number shows each physical memory equipment 15. The address in storage is the address which shows the storage region within physical memory equipment 15.

[0041] Drawing 5 is a flow chart explaining processing actuation of the disk array control device 13 at the time of the disk array control device 13 reading the information in the operation information acquisition section 223 of physical memory equipment 15, and explains this hereafter.

[0042] (1) After the storage subsystem 12 is started, the disk array control device 13 initializes storage occupancy hour entry 14A, and transmits an acquisition demand of the access occupancy hour entry of the physical memory equipment 15 to two or more physical memory

equipments 15 connected after that (steps 371 and 372).

[0043] (2) Next, the disk array control unit 13 stores the access occupancy hour entry of reception and each physical memory equipment in storage occupancy hour entry 14A for an access occupancy hour entry from each physical memory equipment 15 (steps 373 and 374). [0044] In addition, the timing of acquisition of the access occupancy information on the disk array control unit 13 mentioned above. The method which reads the access occupancy information by access to physical memory equipment 23 from the module of others by host I/O, backup, etc. to a fixed time interval from the occupancy time amount total section 227 in each physical memory equipment 15. When an access occupancy hour entry acquisition demand is transmitted to the disk array control unit 13 from other modules (for example, a host 10 and a control terminal 17), it is various and dependent on a design.

[0045] The access occupancy hour entry acquired by the above-mentioned is recorded on the occupancy time amount total table in the disk array control device 13.

[0046] Drawing 6 is a flow chart explaining processing actuation of the physical memory field control device 22 in physical memory equipment 15, next explains this.

[0047] (1) The physical memory field control device 22 receives the disk array configuration information 14 which is information on the logical address and the physical address in the physical memory field 23 of the data 15 transmitted from the disk array control device 13 at the time of starting of the storage subsystem 12, i.e., physical memory equipment, such as correspondence, (steps 400 and 401).

[0048] (2) The physical memory field control unit 22 which received the disk array configuration information 14 performs creation initialization of configuration / classification information management table of configuration / classification Research and Data Processing Department 226 in the operation information acquisition section 223, or the occupancy time amount total table of the occupancy time amount total section 227 based on the information (step 402).

[0049] (3) After initialization processing of the table in step 402 is completed, in order to make this physical memory equipment 15 recognize an accessible thing, notify that initialization processing was completed to the disk array control unit 13. In addition, it is not necessary to transmit the information to which physical disk equipment changed in the accessible condition to the disk array control device 13. When the access directions to a certain storage region come to the physical memory field control unit 22 of physical memory equipment 15 after the fixed time amount, in this case, the physical memory field control unit 22 if it is in a condition accessible to the physical memory field 23 in order to perform predetermined processing. Access the physical memory field 23, perform predetermined processing, and if impossible [whether processing predetermined in the condition of having stored access information in the access request information storage section 225, and having become accessible to the physical memory field is performed, and] or you may make it not receive a certain access directions to a storage region until it comes to resemble the physical memory field 23 in the accessible condition, in order to perform predetermined processing (step 403).

[0050] (4) After that, the physical memory field control device 22 waits to transmit host I/O, and a physical memory equipment operation information acquisition demand instruction or new disk array configuration information from the disk array control device 13, and receives it (step 404).

[0051] (5) At step 404, if host I/O is received from the disk array control unit 13, when the I/O judges lead processing or light processing and it is lead processing, the data-processing control section 224 will confirm whether the data which should be read exist in cache memory 22 (steps 405 and 406).

[0052] (6) When the data is read from cache memory 22 when the data exists in cache memory 22, and the data does not exist in cache memory 22 with the check of step 406, read the data from the physical memory field 23, and transmit data to the disk array control unit 13 (steps 407, 709, and 408).

[0053] (7) At step 405, when judged with host I/O being light processing, the data-processing control section 224 receives the light data transmitted by the host 10, and writes the light data in cache memory 22 (steps 410 and 411).

[0054] (8) And the data-processing control section 224 stores the above-mentioned light data in

the physical memory field 23 while notifying the notice of data write-in termination to the disk array control device 13 (steps 412 and 413).

[0055] The data-processing control section 224 to cache memory 22 after processing of step 408, or processing of step 413 (9) In access **. Or the JOB classification information on the information on whether the physical memory field 23 was accessed, a random lead, a sequential lead, etc.. The access classification of the JOB classification information on the random read/write at the time of writing light data in the physical memory field 23, sequential read/write, etc. is recognized. The occupancy hour entry which accessed cache memory 22 or the physical memory field 23 is stored in the occupancy time amount total section 227 in the operation information acquisition section 223 for every access classification (steps 414 and 415).

[0056] (10) If new disk array configuration information is received from the disk array control device 13 at step 404, the data-processing control section 224 will rewrite the information in configuration / classification Research and Data Processing Department 226, the operation information acquisition section 223, corresponding to new disk array configuration information (steps 418 and 419).

[0057] (11) When a physical memory equipment operation information acquisition demand instruction is received from the disk array control device 13 at step 404, the data-processing control section 224 reads the access occupancy hour entry of the physical memory equipment 15 stored in the occupancy time amount total section 227 in the operation information acquisition section 223, and transmits it to the disk array control device 13 (steps 416 and 417).

[0058] In addition, it may be made to perform transmission to the disk array control device 13 from the physical memory equipment 15 of the occupancy hour entry in processing of step 417 mentioned above to the disk array control device 13 with a fixed time interval automatically from physical memory equipment 15. In this case, said physical memory equipment operation information acquisition demand instruction is not transmitted to physical memory equipment 13 from the disk array control unit 13.

[0059] Drawing 7 is drawing explaining the example of a configuration of the information 141 corresponding to logic/physics in the table for managing correspondence with the address of a logic storage region and the address of a physical memory field which are held in the disk array control unit 13.

[0060] The disk array control unit 13 has managed correspondence with the address of the logic storage region in the physical memory field 23 in two or more physical memory equipments 15 connected, and the address of the physical memory field in the logic storage region. Each of the logic storage region number 500 given to a specific logic storage region, the logical address 510, the physical address 520 by the storage number 521 with a physical storage region with the logic storage region and the address 522 of a physical storage region, the RAID level 530 that shows the engine performance of the physical-memory equipment 15, and the parity group number 540 to which the physical-memory equipment 15 belongs is matched, and the information 141 corresponding to the logic/physics used for this is constituted, as shown in drawing 7. When processing of a lead, a light, etc. specifies the address of a logic storage region from a host 10, a control terminal 17, and other modules (for example, other disk array control units etc.) and it is accessed to the self-disk array control unit 13 by having such information 141 corresponding to logic/physics, the disk array control unit 13 can change the address of a logic field into the address of a physical storage region, and can perform read/write processing of data correctly to physical memory equipment 15.

[0061] Drawing 8 is drawing showing the example of the logic field operating condition information 144 stored in the disk array control unit 13, and the storage occupancy hour entry 141 of physical field operating condition information 145 grade. Such information is constituted as an occupancy time amount total table.

[0062] The disk array control device 13 reads periodically the access occupancy information on the physical memory field 23 of the physical memory equipment 15 by access from the module of others by host I/O, backup, etc. from the occupancy time amount total section 227 in each physical memory equipment 15, and records the access occupancy hour entry on the occupancy time amount total table in the receiving disk array control device 13. The example shown in

drawing 8 is every logic storage region number 601 and I/O. Occupancy time amount is totaled every JOB classification 602. I/O As a JOB classification, although 670 and a total of 680 are shown by the example of illustration at the time of the sequential lead 610, the sequential light data 620, the sequential light parity 630, the random lead 640, the random light parity 660, and a cache hit, it is I/O of further others. There may be JOB classification.

[0063] The occupancy time amount read from physical memory equipment 15 may be record by the accumulation value of not only the above-mentioned but the access occupancy time amount for every I/O, and a universal time amount value and a time amount value peculiar to a machine. Moreover, in the disk array control unit 13, the access occupancy time amount of each logic storage region or a physical memory field may be edited, and an occupancy hour entry table may newly be created based on the value which found the access occupancy time amount for every physical memory equipment and every parity group.

[0064] When a host 10 and control terminal 17 grade give the operation information acquisition demand of physical memory equipment to the disk array subsystem 12 by the above-mentioned, even if a host 10 and control terminal 17 grade access direct physical memory equipment 15 and do not acquire the access occupancy hour entry of a logic storage region or a physical memory field, they become possible [acquiring the access occupancy hour entry of the logic storage region and physical memory field from the disk array control unit 13].

[0065] Drawing 9 is drawing showing the example of a configuration of the table which manages matching with the address of a logic storage region and the address of a physical memory field which are stored in configuration / classification Research and Data Processing Department 226 in the operation information acquisition section 223 in physical memory equipment 15.

[0066] From the disk array control device 13, when the relation of the address of a logic storage region and the address of a physical-memory field which are produced by the time of starting of migration of a logic storage region, etc. changes, the data-processing control section 224 in physical-memory equipment 15 receives the matching information on the logic storage region and physical-memory field, and stores it in a correspondence table as shown in configuration / classification Research and Data Processing Department 226 at drawing 9. This correspondence table is constituted by physical ADORE 7520 by the logic storage region number 700 given to a specific logic storage region, the logical address 710, and the storage number 721 with a physical storage region with that logic storage region and the address 722 of a physical storage region. Thereby, it can recognize which address logic storage region physical memory equipment 15 has where of the physical memory regional address in self-physical memory equipment 15.

[0067] Drawing 10 is drawing showing the example of a configuration of the table of the occupancy hour entry by access to the storage region by I/O accumulated and stored at the occupancy time amount total section 227 of the operation information acquisition section 223 in physical memory equipment 15.

[0068] This table is every logic storage region number 801 and I/O. Occupancy time amount is totaled every JOB classification 802. I/O As a JOB classification, in the example of illustration, although 870 and a total of 880 are shown at the time of the sequential lead 810, the sequential light data 820, the sequential light parity 830, the random lead 840, the random light parity 860, and a cache hit, there may be I/OJOB classification of further others.

[0069] If the data-processing control section 224 in physical memory equipment 15 has access in a store by I/O from a host etc., it will accumulate the occupancy time amount 890 in the occupancy time amount total section 227 every JOB classification 802 of access about each logic storage region 801 with access. This becomes possible to obtain the relation between the number of the logic storage region in physical memory equipment 15, the occupancy time amount by the access classification to the logic storage region, and the sum total occupancy time amount within a certain time amount within physical memory equipment 15.

[0070] According to the operation gestalt of this invention mentioned above, the occupancy hour entry of the storage for every I/O can be acquired, and acquisition of the occupancy hour entry of the storage by access to a storage region can be realized within physical memory equipment.

[0071] According to the operation gestalt of this invention, the thing of two or more physical

memory equipments which constitute a storage subsystem for which the occupancy time amount of the logic storage region of physical memory equipment is acquired becomes respectively possible by the above-mentioned, and the thing of each logic storage region to do for occupancy time amount (real operating time) acquisition becomes possible by the system configuration unacquirable [with a disk array control unit]. Moreover, in order according to the operation gestalt of this invention mentioned above to be able to make small the error of the analysis of the utilization factor of the logic storage region of said storage subsystem, or utilization factor prediction and to perform optimum performance tuning more by taking into consideration the effect on the physical memory equipment for every I/O, the access occupancy hour entry for every I/O to physical memory equipment with a more high precision is acquirable.

[0072]

[Effect of the Invention] As explained above, according to this invention, physical memory equipment can acquire the occupancy time amount of a logic storage region, and the occupancy time amount (real operating time) of each logic storage region can be acquired by the system configuration unacquirable [with a disk array control unit].

[0073] Moreover, according to this invention, since the access occupancy hour entry for every I/O to physical memory equipment is acquirable, it becomes possible to perform the analysis of the utilization factor of the logic storage region of a storage subsystem and the prediction of a utilization factor in consideration of the effect on the physical memory equipment for every I/O with a small error, and the engine performance of the more nearly optimal storage subsystem can be tuned up.

[Translation done.]

* NOTICES *

- JP0 and NCIP1 are not responsible for any damages caused by the use of this translation.
- 1.This document has been translated by computer. So the translation may not reflect the original precisely.
- 2.*** shows the word which can not be translated.
- 3.In the drawings, any words are not translated.

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]

- [Drawing 1] It is the block diagram showing the configuration of the computer system equipped with the storage subsystem by this invention.
 - [Drawing 2] It is the block diagram showing the configuration of physical memory equipment.
 - [Drawing 3] It is a flow chart explaining processing actuation of a disk array control device when a storage subsystem is started.
 - [Drawing 4] It is a flow chart explaining processing actuation of a disk array control device when correspondence of the address of a logic storage region and the address of a physical memory field changes.
 - [Drawing 5] It is a flow chart explaining processing actuation of the disk array control device at the time of a disk array control device reading the information on operation information acquisition circles of physical memory equipment.
 - [Drawing 6] It is a flow chart explaining processing actuation of the physical memory field control device in physical memory equipment.
 - [Drawing 7] It is drawing explaining the example of a configuration of the information corresponding to the logic/physics which manages correspondence of the address of the logic storage region and physical memory field which are held in the disk array control unit.
 - [Drawing 8] It is drawing showing the example of storage occupancy hour entries, such as logic field operating condition information stored in a disk array control unit, and physical field operating condition information.
 - [Drawing 9] It is drawing showing the example of a configuration of the table which manages matching of the address of the logic storage region and physical memory field which are stored in configuration / classification Research and Data Processing Department in physical memory equipment.
 - [Drawing 10] It is drawing showing the example of a configuration of the table of the occupancy hour entry by access to the storage region accumulated and stored at the occupancy time amount total section of the operation information acquisition section in physical memory equipment.
- [Description of Notations]
- 10 Host
 - 12 Storage Subsystem
 - 13 Disk Array Control Unit
 - 14 Disk Array Control Information
 - 15 Physical Memory Equipment
 - 16 Disk Array
 - 17 Control Terminal
 - 18 I/O Bus
 - 19 Network
 - 22 Physical Memory Field Control Unit
 - 23 Physical Memory Field
 - 130 Read/write Processing Section

- 131 Operating Condition Information Acquisition Processing Section
- 132 Relocation Decision Processing Section
- 133 Relocation Executive Operation Section
- 141 Information corresponding to Logic/Physics
- 142 Class Configuration Information
- 143 Class Attribute Information
- 144 Logic Field Operating Condition Information
- 145 Physical Field Operating Condition Information
- 146 Relocation Decision Horizon Information
- 147 Relocation Activation Time Information
- 148 Free-Space Information
- 149 Relocation Information
- 14A Storage occupancy hour entry
- 221 Physical Memory Field Control Section
- 222 Cache Memory
- 223 Operation Information Acquisition Section
- 224 Data-Processing Control Section
- 225 Access Request Information Storage Section
- 226 Configuration / Classification Research and Data Processing Department
- 227 Occupancy Time Amount Total Section

[Translation done.]

(19)日本特許庁(JP) (12)公開特許公報(A)

(11)特許出願公開番号
特開2002-157091
(P2002-157091A)
(43)公開日 平成14年5月31日(2002.5.31)

(51)Int.Cl. ⁷	G 0 6 F 3/06	F 1	G 0 6 F 3/06	チーゴット(参考)
識別記号	3 0 2		3 0 2 A	5 B 0 1 8
	3 0 1		3 0 1 X	5 B 0 6 5
	3 0 4		3 0 4 N	
	5 4 0		5 4 0	
	3 2 0		3 2 0 L	
	12/16		12/16	
審査請求	未請求	請求項の数	5	OL (全 16 頁)

(21)出願番号	特開2000-353010(P2000-353010)	(71)出願人	000005108 株式会社日立製作所
(22)出願日	平成12年11月20日(2000.11.20)	(72)発明者	江口 賢裕 株式会社日立製作所 東京都千代田区神田墨田四丁目6番地
		(73)発明者	茂木 和彦 株式会社日立製作所システム開発研究所内 神奈川県横浜市磯城区王禅寺1099番地 株
		(74)代理人	井理士 鈴木 市郎 (外1名) 株式会社日立製作所システム開発研究所内 100093492 神奈川県横浜市磯城区王禅寺1099番地 株

最終頁に続く

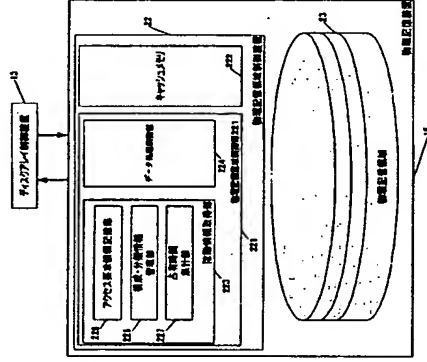
(54)【発明の名称】 ストレージサブシステム及びそのシステムに使用する記憶装置

(57)【要約】

【課題】 物理記憶装置において、物理記憶装置の占有時間を取得し、より精度の高い物理記憶装置への1/O毎のアクセス占有時間情報を取得する。

【解決手段】 個々の物理記憶装置15上にある物理記憶領域制御装置22に、ホストからのアクセス要求に関する情報を記憶するテーブル225と、アクセスに関する情報を集計するテーブル227と、アクセス要求の構成を分類する管理情報のテーブル226と、ディスクアレイ制御装置13から物理記憶領域の構成情報や分類情報を取得し、必要に応じて物理記憶領域の構成情報や分類情報をディスクアレイ制御装置にリクエストするデータ処理制御部224を備える。また、ディスクアレイ制御装置13に、物理記憶装置上にある物理記憶領域からのリクエストに応じて、現在のディスクアレイの構成情報を物理記憶領域制御装置に送信する手段を設ける。

図2



【特許請求の範囲】

【請求項1】 1または複数の計算機に接続され、複数の物理記憶装置と、これらの複数の物理記憶装置の使用状況情報を取得する手段と、前記計算機がリード/ライト対象とする物理記憶領域と前記物理記憶装置の物理記憶領域との対応付けを行う手段とを有するストレージサブシステムにおいて、前記複数の物理記憶装置のそれぞれは、物理記憶領域制御装置を備え、該物理記憶領域制御装置は、物理記憶領域の使用状況を取得する手段を有することを特徴とするストレージサブシステム。

【請求項2】 1または複数の計算機に接続され、複数の物理記憶装置と、これらの複数の物理記憶装置の使用状況情報を取得する手段と、前記計算機がリード/ライト対象とする物理記憶領域と前記物理記憶装置の物理記憶領域との対応付けを行う手段とを有するストレージサブシステムにおいて、前記複数の物理記憶装置のそれぞれは、物理記憶領域制御装置を備え、該物理記憶領域制御装置は、物理記憶領域の使用状況を取得する手段を有することを特徴とするストレージサブシステム。

【請求項3】 ストレージサブシステムを構成する物理記憶装置において、物理記憶領域制御装置を備え、該物理記憶領域制御装置は、物理記憶領域の使用状況情報を取得する手段を有することを特徴とする請求項1または2記載のストレージサブシステムに使用する物理記憶装置。

【請求項4】 前記物理記憶領域制御装置は、取得した物理記憶領域の使用状況情報を格納する手段をさらに備えることを特徴とする請求項3記載の物理記憶装置。

【請求項5】 前記物理記憶領域制御装置は、前記制御装置により受信する物理記憶領域の使用状況情報の取得状況を物理記憶領域に送信する手段と、前記制御装置より受信する物理記憶領域の使用状況情報と物理記憶領域との対応付けした情報を格納する手段をさらに備えることを特徴とする請求項3または4記載の物理記憶装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 本発明は、ストレージサブシステム及びそのシステムに使用する記憶装置に関する。【0002】

【従来の技術】 コンピュータシステムに使用される高性能な二次記憶システムとして、ディスクアレイシステムが知られている。

【0003】 ディスクアレイシステムは、複数の物理記憶装置をアレイ状に配置し、各物理記憶装置にデータを分割して格納しておき、前記各物理記憶装置を個別に動作させて、前記各物理記憶装置に分割して格納されたデータのリード/ライトを高速度に行うことを可能としたシステムである。

【0004】 ディスクアレイシステムに関する従来技術として、例えば、D.A.Patterson, G.Gibson, and R.H.Klks, "A Case for Redundant Arrays of Inexpensive Disks (ディスクアレイ)" (in Proc. ACM SIGMOD, pp. 109-116, June 1988) 等に記載された技術が知られている。この従来技術は、冗長性を付加したディスクアレイシステムに対し、その構成に応じてレベル1からレベル5の種別を与えておくというものである。また、これらの種別に、冗長性無しのディスクアレイシステムを加え、これをレベル0と呼ぶこともある。前述の各レベルは、冗長性等に応じて異なる構成として実現するため、コストや性能特性等が異なる。そして、ディスクアレイシステムを構築するにあたって複数のレベルのアレイ(物理記憶装置の組)を混在させることも多い。ここで、冗長性を付加したディスクアレイの組をパリティグループと呼ぶ。また、物理記憶装置についても性能や容量等によりコストが異なり、ディスクアレイシステムを構築するにあたって異質なコストパフォーマンスを実現するために、やはり性能や容量の異なる複数の物理記憶装置を用いることがある。

【0005】 ディスクアレイシステムに格納されるデータは、前述のような前記物理記憶装置に分散して配置される。このため、ディスクアレイシステムは、ディスクアレイシステムに接続されるホストコンピュータがアクセスする物理記憶領域と前記物理記憶装置の記憶領域をセグメントに分割し、すなわち、アドレス変換を行う必要がある。

【0006】 アドレス変換を行うディスクアレイシステムに関する従来技術として、例えば、特開平9-274544号公報等に記載された技術が知られている。この従来技術は、ホストコンピュータからの物理記憶領域に対する1/Oアクセスについての情報を取得する手段と、物理記憶領域の物理記憶領域への対応付けを要として物理記憶領域を行う手段とにより、格納されたデータの物理記憶領域を実現するというものであり、この公報には、物理記憶領域への1/Oアクセス占有時間情報を、ディスクアレイ制御装置が取得する技術が示されている。

【0007】 物理記憶領域への1/Oアクセス占有時間情報は、ディスクアレイシステムの負荷分散を行うため、物理記憶領域の物理記憶領域への対応付けを変更し

て物理的再配置を行う際に、元になるデータとなるため重要である。

【発明が解決しようとする課題】前述した公報に記載された従来技術は、論理記憶領域への1/0による占有時間をディスクアレイ制御装置が取得するというものであるが、このディスクアレイ制御装置に示された方法は、次に説明するような問題点を有している。

【0009】まず、ある論理記憶領域にデータの書き込み(ライト)が行われた場合を考える。この場合、当該論理記憶領域に対応する物理記憶領域にデータのライトが行われる。物理記憶領域は、物理記憶装置内に有り、物理記憶装置は、主に物理記憶領域制御部、データをキャッシュするキャッシュメモリ及び物理記憶領域により構成される。そして、データを物理記憶領域にライトする場合、物理記憶領域制御部がライトデータをキャッシュに蓄込んだ時点で書き込み終了の応答がディスクアレイ制御装置に通知される。このため、前述した従来技術は、実際にデータをライトするための物理記憶領域にアクセスした時間が利らないという問題点を有することになる。

【0010】次に、ある論理記憶領域にデータの読み込み(リード)が行われた場合を考える。この場合、当該論理記憶領域に対応する物理記憶領域にあるデータのリードが行われるが、実際には、物理記憶装置内のキャッシュメモリにそのリードデータがあった場合、物理記憶領域にはアクセスせず、キャッシュメモリにアクセスし、そのデータを返す。このため、前述した従来技術は、データリードのために実際に物理記憶装置にはアクセスがあったかどうかを判断することができず、また、物理記憶領域にアクセスした正確な時間が利かないという問題点を有することになる。

【0011】また、A、B、C、Dと複数回にわたってある論理記憶領域にアクセスがあった場合を考える。そして、Aの応答をA'、Bの応答をB'等とし、アクセスAの時刻をA(t)、アクセスBの時刻をB(t)、応答A'の時刻をA'(t)等とする。ここで、アクセスAは、データリードで物理記憶装置にアクセスしてデータをリードし、アクセスBは、データリードでデータがキャッシュにあるため物理記憶装置にはアクセスしないのでその応答B'があったものと仮定する。この場合、先にあったアクセスAの応答A'よりも後からきたアクセスBの応答B'の方が先になる。すなわち、A(t)<B(t)、A'(t)>B'(t)となる。このとき、論理記憶領域への1/0による占有時間をディスクアレイ制御装置において取得する従来技術は、どの1/0が物理記憶装置内の物理記憶領域にどのくらいアクセスしたか、あるいは、物理記憶装置のキャッシュにヒットしたか否かをしることができきないという問題点を生じ

る。

【0012】本発明の目的は、前述した従来技術の問題点を解決し、物理記憶装置で論理記憶領域の占有時間を取得することによって、ディスクアレイ制御装置のみで取得不可能なシステム構成で各論理記憶領域の占有時間(裏書き時間)を取得できるようにしたストレージシステムを提供し、かつ、これに使用する記憶装置を提供することにある。

【0013】また、本発明の目的は、1つの1/0毎の物理記憶装置への影響を考慮することによって、前記ディスクアレイシステムの論理記憶領域の利用率の解析や利用率予測の誤差を小さくすることができ、より最適な性能チューニングを行うために、より精度の高い物理記憶装置への1/0毎のアクセス占有時間情報を取得することができるとするストレージサブシステムを提供し、かつ、これに使用する記憶装置を提供することにある。

【0014】
【課題を解決するための手段】本発明によれば前記目的は、1または複数の計算機に接続され、複数の物理記憶装置と、これらの複数の物理記憶装置の使用状況情報を取得する手段と、前記計算機がリード/ライト対象とする論理記憶領域と前記物理記憶装置の物理記憶領域との対応付けを行う手段とを有するストレージサブシステムにおいて、前記複数の物理記憶装置の使用状況情報を取得する手段と、前記計算機がリード/ライト対象とする論理記憶領域と前記物理記憶装置の物理記憶領域の対応付けを行う手段とが、前述複数の物理記憶装置を制御する制御装置内に設けられ、前記制御装置が、さらに、物理記憶装置の論理記憶領域と物理記憶装置の物理記憶領域との対応付けを行った情報を前記複数のそれぞれの物理記憶装置に送信する手段を備え、前記複数の物理記憶装置のそれぞれの物理記憶装置が、物理記憶領域制御装置を備え、該物理記憶装置が、物理記憶領域制御装置の使用状況を取得する手段を有することにより達成される。

【0015】また、前記目的は、ストレージサブシステムを構成する物理記憶装置において、物理記憶領域制御装置を備え、該物理記憶領域制御装置が、物理記憶領域の使用状況を取得する手段、取得した物理記憶領域の使用状況情報を格納する手段、前記物理記憶装置より受得する物理記憶領域の使用状況情報の取得要求に応じて、自物理記憶装置の物理記憶領域の使用状況情報を前記制御装置に送信する手段、及び、前記制御装置より受得する物理記憶装置の論理記憶領域と物理記憶領域とを対応付けた情報を格納する手段を備えることにより達成される。

【0016】
【発明の実施の形態】以下、本発明によるストレージサブシステム及びそのシステムに使用する記憶装置の実施形態を図面により詳細に説明する。

【0017】図1は本発明によるストレージサブシステムを備えた計算機システムの構成を示すブロック図、図

50

2は物理記憶装置の構成を示すブロック図である。図1、図2において、10はホスト、12はストレージサブシステム、13はディスクアレイ制御装置、14はディスクアレイ制御情報、15は物理記憶装置、16はディスクアレイ、17は制御端末、18は1/0バス、19はネットワーク、22は物理記憶領域制御装置、23は物理記憶領域、130はリード/ライト処理部、131は使用状況情報取得処理部、132は再配置判断処理部、133は再配置実行処理部、141は論理/物理対応情報、142はクラス構成情報、143はクラス属性情報、144は論理領域使用状況情報、145は物理領域使用状況情報、146は再配置判断対称時間情報、147は再配置実行時刻情報、148は未使用領域情報、149は再配置情報、14Aは記憶装置占有時間情報、221は物理記憶領域制御部、222はキャッシュメモリ、223は移動情報取得部、224はデータ処理制御部、225はアクセス要求情報記憶部、226は構成・分類情報管理部、227は占有時間累計部である。

【0018】図1に示す計算機システムは、上位の計算機である1または複数のホスト10、ストレージサブシステム12、制御端末17から構成される。ホスト10は、ストレージサブシステム12に1/0バス18で接続され、ストレージサブシステム12に対してデータのリードやライト処理のための1/0を発行する。この1/0を行う際、ホスト10は、ストレージサブシステム12の論理的な記憶領域を指定する。すなわち、ホスト10は、ストレージサブシステム12の論理的な記憶領域のアドレスによりアクセスを行う。また、1/0バス18は、例えば、ESCON、SCSI、ファイバチャネル等により構成される。

【0019】ストレージサブシステム12は、ディスクアレイ制御装置13及び複数の物理記憶装置15から構成される。ディスクアレイ制御装置13は、リード/ライト処理部130と、使用状況情報取得処理部131と、再配置判断処理部132と、再配置実行処理部133とを備え、これらの処理部が、リード/ライト処理、使用状況情報取得処理、再配置判断処理、再配置実行処理等の処理を行う。また、ストレージサブシステム12のディスクアレイ制御部は、論理記憶領域/物理記憶領域対応情報141、クラス構成情報142、クラス属性情報143等のディスクアレイ構成情報1400と、論理領域使用状況情報144、物理領域使用状況情報145等の記憶装置占有時間情報14Aと、再配置判断対称時間情報146と、再配置実行時刻情報147と、未使用領域情報148と、再配置情報149等を保持している。なお、前述したディスクアレイ構成情報14には、前述した情報の他、パーティション情報やRAIDレベル情報等が含まれてもよい。

【0020】また、ホスト10、ディスクアレイ制御装置13及び制御端末17は、相互にネットワーク19に

50

より接続されている。ネットワーク19は、例えば、イーサネット(登録商標)、FDDI、ファイバチャネル等により構成されてよい。制御端末17は、通常、ストレージサブシステム12の保守・管理等を行うために使用される。

【0021】また、ホスト10、ディスクアレイ制御装置13及び制御端末17には、それぞれでの処理を行うためのメモリ、CPU等の計算機において必ず存在する構成要素をそれぞれ存在するが、本発明の実施形態の説明においては重要であるため、ここでは明記しない。

【0022】前述のストレージサブシステム12内に設けられる複数の物理記憶装置15は、物理記憶装置の性能毎にクラス分けされて、クラス毎にディスクアレイ16を構成している。また、ここでは、明示的に示していないが、複数の物理記憶装置を使用して、パーティションが構成されている。そして、物理記憶装置15のそれぞれは、図2に示すように、物理記憶領域23とこの物理記憶領域23を制御する物理記憶領域制御装置22とにより構成され、物理記憶領域23には、様々なデータが格納されている。

【0023】また、前述したように、ホスト10からは物理記憶領域23のアドレスは直接見えておらず、ホスト10は、複数の物理記憶領域23上にある複数の論理的な記憶領域にあるデータにアクセスを行う。すなわち、ホスト10は、ストレージサブシステム12内の各物理記憶装置15の記憶領域にあるデータに対して、論理記憶領域を指定してアクセスを行う。

【0024】ディスクアレイ制御装置13は、複数の物理記憶装置15と接続されており、複数の物理記憶装置15を制御したり、前記ホスト10から送られてくるリードやライト処理命令1/0を、指定のデータが存在する論理記憶領域のアドレスとその論理記憶領域のアドレスがある物理記憶領域のアドレスとを対応させて、適当な物理記憶装置15にデータ1/0を送信し、ライト処理であれば、ホスト10から送信されてくるデータを物理記憶装置15に送信し、リード処理であれば物理記憶装置15から送信されてくるデータを受信してホスト10に送信する等の処理を行っている。

【0025】物理記憶装置15内に備えられる物理記憶領域制御装置22は、物理記憶領域制御部221とキャッシュメモリ222とにより構成されている。キャッシュメモリ222は、物理記憶領域23に比べてデータのリード/ライトの処理の速度が速い。そして、キャッシュメモリ222は、ディスクアレイ制御装置13から送信されてくるリードまたはライト命令に関するデータに関して決よように使用される。すなわち、ライト処理の場合、ディスクアレイ制御装置13から送信されてくるデータがキャッシュメモリ222にも書き込まれる。また、リード処理の場合、物理記憶領域23からデータ読

50

み出される際に、読み出されたデータは、キャッシュメモリ222に書き込まれる。あるいは、以前のリード処理によって同一のデータがキャッシュメモリ221に有り、そのデータに対してディスクアレイ制御装置13からリード命令として物理記憶装置にきた場合に、そのデータを物理記憶領域23から読み込みます、キャッシュメモリ222に書き込む。これにより、物理記憶装置15の処理性能を上げることができ。

【0026】物理記憶制御部221は、主に発動情報取得部223とデータ処理制御部224とを備えて構成されている。データ処理制御部224は、ディスクアレイ制御装置13から送られてくるデータのリードまたはライト命令を受信する。そして、データ処理制御部224は、受信した命令がリード命令であった場合、キャッシュメモリ222にアクセスし、そのリードデータがキャッシュメモリ222に存在すれば、キャッシュメモリ222からそのリードデータを読み出し、キャッシュメモリ222にそのデータがなければ、物理記憶領域23にアクセスしてそのリードデータを読み出し、ディスクアレイ制御装置13にデータを送信する。また、データ処理制御部224は、受信した命令がライト処理命令であった場合、ディスクアレイ制御装置13から送られてくるデータをキャッシュメモリ222に書き込むと同時に、またはその後で、そのデータを物理記憶領域23に書き込む。ライトデータは、キャッシュメモリ222に書き込みます、直接物理記憶領域23に書き込む。

【0027】発動情報取得部223は、アクセス要求情報記憶部225、構成・分類情報管理部226、占有時間集計部227等により構成される。前述のデータ処理制御部224は、1/0処理によって指定されたデータのある物理記憶領域、そのデータが存在する物理記憶領域23、あるいは、キャッシュメモリ222にアクセスしたときに、そのアクセスの時間情報を1/0処理の処理種別毎（ランダムアクセス、ランダムアレイ制御装置13よりディスクアレイ内の物理記憶領域のアドレスと物理記憶領域23のアドレスとの対応情報や物理記憶装置15の性能等の情報を受信し、構成・分類情報管理部226に記録する。さらに、データ処理制御部224は、複数の1/0処理の時間情報を13等から送られてくるデータのリードまたはライト命令データ等を受信し、アクセス要求情報記憶部225にその命令データを記録する。

【0028】物理記憶装置15は、前述したような構成を有することにより、物理記憶装置15内で1/0処理による物理記憶領域、物理記憶領域23、あるいは、キャッシュメモリ222にアクセスしたときのアクセス時

の時間情報を1/0処理の処理種別毎（ランダムアクセス、ランダムアレイ制御装置13よりディスクアレイ内の物理記憶領域のアドレスと物理記憶領域23のアドレスとの対応情報や物理記憶装置15の性能等の情報を受信し、構成・分類情報管理部226に記録する。さらに、データ処理制御部224は、複数の1/0処理の時間情報を13等から送られてくるデータのリードまたはライト命令データ等を受信し、アクセス要求情報記憶部225にその命令データを記録する。

【0029】図3はストレージサブシステムが起動されたとときのディスクアレイ制御装置の処理動作を説明するフローチャートであり、以下、これについて説明する。【0030】（1）ストレージサブシステム12の発動時、ディスクアレイ制御装置13は、自装置13と接続されている物理記憶装置15に対して、物理記憶装置15内の物理記憶領域23にある物理記憶領域のアドレスと実際の物理記憶領域が存在する物理記憶領域のアドレスとの対応付け情報である物理／物理対応情報142、クラス属性情報143等のディスクアレイ構成情報14を送信する（ステップ300、310）。

【0031】（2）次に、ディスクアレイ制御装置13は、前述した情報の送信により、物理記憶装置15がアクセス可能となったときに、物理記憶領域22から送られてくる物理記憶装置15がアクセス可能なレディ状態に遷移した通知を受信する。このとき、物理記憶装置15は、ディスクアレイ構成情報14による初期化終了の状態となっている（ステップ320）。【0032】（3）続いて、ディスクアレイ制御装置13は、1/0パス18経由でホスト10よりストレージサブシステム12に、そのストレージサブシステム12内の物理記憶領域に対してリード／ライト処理のホスト1/0を送信してきたものや、ディスクアレイ制御装置同士で命令やデータを受け渡す等の様々なデータを受信する（ステップ330）。

【0033】（4）前記受信データとしてホスト1/0を受信した場合、ディスクアレイ制御装置13は、ホスト1/0により指定された物理記憶領域に対するリードまたはライト要求を受信し、その物理記憶領域のアドレス（物理アドレス）を物理記憶領域のアドレス（物理アドレス）に変換する物理／物理対応情報141を用いて、その物理記憶領域23にアクセスする物理記憶領域23のアドレスを求め（ステップ340、350）。

【0034】（5）ディスクアレイ制御装置13は、所定のデータが存在する物理記憶領域のアドレスを指定し、リード処理の場合、前述の物理アドレスを有する物理記憶装置からリードデータを搬出し、ホスト10にリードデータを搬送し、ライト処理の場合、ホスト10から搬送されたライトデータを受信し、その物理アドレスを持つ物理記憶装置にライトデータを送信する（ステップ360）。

【0035】図4は物理記憶領域のアドレスと物理記憶

領域のアドレスの対応が変化した場合のディスクアレイ制御装置の処理動作を説明するフローチャートであり、以下、これについて説明する。

【0036】（1）ディスクアレイ制御装置13は、物理記憶装置15の増設やRAIDレベルの変化、物理記憶領域が存在する物理記憶領域アドレスとは別の物理記憶領域のアドレスに移動する等によって、物理記憶領域のアドレスと物理記憶領域のアドレスの対応が変化したことを監視し、変化があった場合、再度、物理記憶装置15に対して、物理記憶装置15内の物理記憶領域23にある物理記憶領域のアドレスと実際の物理記憶領域が存在する物理記憶領域のアドレスとの対応付け情報である物理／物理対応情報141、クラス構成情報142、クラス属性情報143等のディスクアレイ構成情報14を送信する（ステップ310）。

【0037】（2）次に、ディスクアレイ制御装置13は、前述した情報の送信により、物理記憶装置15がアクセス可能となったときに、物理記憶領域22から送られてくる物理記憶装置15がアクセス可能なレディ状態に遷移した通知を受信する。このとき、物理記憶装置15は、ディスクアレイ構成情報14による更新終了の状態となっている（ステップ320）。

【0038】（3）その後の処理は、図3により説明したステップ330、340、350、360の場合と同様に実行される（ステップ330、340、350、360）。

【0039】なお、前述したステップ320、3201において、ディスクアレイ制御装置13は、物理記憶領域22から、物理記憶装置15がアクセス可能な状態に遷移したという情報を受信しなくてもよい。この場合、ある定まった時間後に物理記憶装置15に対してリード／ライト等のアクセス可能な状態になっていると仮定し、何らかの記憶領域へのアクセス指示がディスクアレイ制御装置13にきた場合に、物理記憶装置15に所定の処理を行うためのアクセス処理を行えばよい。また、所定のアクセスに対する応答がない場合、再度、所定の処理を行うためのアクセス処理を行うか、あるいは、応答が帰ってくるまで待ち、一定時間中に応答がない場合、何らかの記憶領域へのアクセス指示を出したモジュールに対してその旨を伝える方式としてもよい。

【0040】また、前述における物理／物理対応情報141は、物理記憶領域と物理記憶領域とを対応させる情報である。そして、物理アドレスは、ホスト10が前記リード／ライト処理130で用いる物理記憶領域を示すアドレスである。また、物理アドレスは、実際にデータが格納される物理記憶装置15上の領域を示すアドレスであり、物理記憶装置番号及び物理記憶装置内アドレスからなる。記憶装置番号は、個々の物理記憶装置15を示す。記憶装置内アドレスは、物理記憶装置15内の記憶領域を示すアドレスである。

【0041】図5はディスクアレイ制御装置13が物理記憶装置15の発動情報取得部223の動作を説明するフローチャートであり、以下、これについて説明する。

【0042】（1）ディスクアレイ制御装置13は、ストレージサブシステム12が起動された後、物理記憶装置14Aを初期化し、その後、接続されている複数の物理記憶装置15にその物理記憶装置15のアクセス占有時間情報の取得要求を送信する（ステップ371、372）。

【0043】（2）次に、ディスクアレイ制御装置13は、各物理記憶装置15よりアクセス占有時間情報を受取り、各物理記憶装置のアクセス占有時間情報を記憶装置占有時間情報14Aに格納する（ステップ373、374）。

【0044】なお、前述したディスクアレイ制御装置13のアクセス占有情報の取得のタイミングは、ホスト1/0やバックアップ等による他のモジュールから物理記憶装置23に対するアクセスによるアクセス占有情報を各物理記憶装置15内の占有時間集計部227から一定時間間隔に読み出す方式や、他のモジュール（例えばホスト10や制御端末17）からアクセス占有時間情報取得要求がディスクアレイ制御装置13に送信された際等、様々であり、設計に依存する。

【0045】前述により取得されたアクセス占有時間情報は、ディスクアレイ制御装置13内の占有時間集計テーブルに記録される。

【0046】図6は物理記憶装置15内の物理記憶領域22の処理動作を説明するフローチャートであり、次に、これについて説明する。

【0047】（1）物理記憶装置15は、ストレージサブシステム12の発動時にディスクアレイ制御装置13より送られてくるデータ、すなわち、物理記憶装置15の物理記憶領域23にある物理アドレスと物理アドレスとの対応等の情報であるディスクアレイ構成情報14を受信する（ステップ400、401）。

【0048】（2）ディスクアレイ構成情報14を受信した物理記憶装置15は、その情報を元に、稼働情報取得部223の構成・分類情報管理部226の構成・分類情報管理テーブルや占有時間集計部227の占有時間集計テーブルの作成初期化を行う（ステップ402）。

【0049】（3）ステップ402でのテーブルの初期化処理が終了すると、この物理記憶装置15にアクセス可能であることを認識させるためディスクアレイ制御装置13に初期化処理が終了したことを通知する。なお、ディスクアレイ制御装置13に物理アドレスがアクセス可能な状態に遷移した情報を送信しなくてもよい。この場合、定まった時間後に物理記憶装置15の物理記

物理記憶制御装置 2 2 に対して何らかの記憶領域へのアクセス指示が行われる場合、物理記憶制御装置 2 2 は、所定の処理を行うために物理記憶装置 2 3 にアクセスし、所定の状態であれば、その物理記憶装置 2 3 にアクセスし、所定の処理を行い、不可能であれば、アクセス要求情報配記部 2 2 5 にアクセスし、物理記憶装置 2 3 にアクセスし、所定の状態になった状態で所定の処理を行うか、あるいは、所定の処理を行うために物理記憶装置 2 3 にアクセス可能な状態になるまで、記憶領域への何らかのアクセスを受け付けられないようにしてもよい。(ステップ 4 0 3)。

【0050】(4) その後、物理記憶領域制御装置22は、ディスクアレイ制御装置13からホスト10や、物理記憶装置移動情報取得要求命令、あるいは、新たなディスクアレイ構成情報が送信されてくるのを待つてそれを受信する(ステップ404)。

【0051】(5) ステップ404で、ディスプレイレイアウト装置13からホスト1/Oを受信すると、その1/Oがリード処理かライト処理かを判定し、リード処理であった場合、データ処理制御部224は、読み出すべきデータがキャッシュメモリ22内に存在するか否かをチェックする(ステップ405、406)。

【0052】(6) ステップ406のチェックで、そのデータがキャッシュメモリ22内に存在した場合、そのデータはキャッシュメモリ22から読み出し、また、そのデータがキャッシュメモリ22内に存在しなかった場合、物理記憶領域23からそのデータを読み出し、データをディスクアレイ制御装置13に転送する(ステップ407、709、408)。

【0053】(7) ステップ405で、ホスト1/0が
ライト処理であると判定された場合、データ処理制御部
224は、ホスト10から転送されたライトデータを受
信し、キャッシュメモリ2にそのライトデータを書き
込む(ステップ410、411)。

【0054】(8) として、データ処理制御部224は、データ書き込み終了通知をディスプレイ制御装置113に通知すると共に、前述のライトデータを物理記憶領域23に格納する(ステップ412、413)。

【0055】(9) ステップ408の処理後、または、ステップ413の処理後、データ制御部224は、キヤッシュメモリ22にアクセスした、あるいは、物理記憶領域23にアクセスしたかの情報、ランダムリード/ライトケーンシヤリールIDなどのJOB種別情報、ライト/ライトにシヤークンシヤリールID/リード/リードのJOB種別情報のアクセス履歴を記録し、アクセス履歴0に、キヤッシュメモリ22にあるいは物理記憶領域23にアクセスした占用時情報を格納する(ステップ23内の15)。

【0062】ディスクドライブ制御装置13は、ホスト1のI/Oバスにアクセス等によるその他のモジュールからのアクセスによる物理記憶装置15の物理記憶領域23へのアクセスと占有権を各物理記憶装置15内の占有時間累計部227から定期的に検出して、そのアクセス占有時間情報を受信ディスクアクセス制御装置13内の占有時間情報データベース11に記録する。図8に示す例は、物理記憶装置601毎、1/O JOB種別602毎に占有時間と、図示例では、シーケンシャルジョブ610、シークジョブ620、ランダムジョブ640、ランダムリプレイ630、ランダムリード640、ランダムリプレイ660、キャッシュヒット670、合計680が示されているが、さらに他の1/O JOB種別があってもよい。

【0063】物理記憶装置15から読み出される占有時間、は、前述に限らず、1/10毎のアクセス占有時間の累積値や、エニバーサル時間間、マシン固有の時間間延び等の記録であってもよい。また、ディッシュアレイ制御装置13において、各物理記憶装置毎や物理記憶領域のアクセス占有時間を構築し、物理記憶装置毎やパーティション毎のアクセス占有時間を求め、物理記憶装置15に基づいて新たな占有時間情報テーブルを作成してもよい。

【0064】前述により、ホスト10や制御部末17等が物理記憶装置の発行情報取得要求をディスクアレキサシステム112に出した場合は、ホスト10や制御部末17等は、直接物理記憶装置15にアクセスして論理記憶7等は、物理記憶装置15に占有時間情報取得しな領域や物理記憶装置のアクセス占有時間情報取得しなくてもディスクアレキサ制御装置13からその論理記憶領域や物理記憶装置のアクセス占有時間情報取得することとなる。

【0065】図9は物理記憶装置15内の移動情報取得部223内の構成・分取情報管理部226に格納される論理記憶領域のアドレスと物理記憶領域のアドレスとの対応付けを管理するテーブルの構成例を示す図である。

【0066】物理記憶装置15内のデータ処理制御部22は、ディスク制御部13より、ストレージシステム2の動作や、物理記憶装置の増減やレイアウトの変更等に応じて生じる管理記憶領域のアドレスと物理記憶領域のアドレスとの対応情報を受け取り、それを構成・分類情報管理テーブル26に図9に示すような対応テーブルに格納する。この対応テーブルは、特定の管理記憶領域に対して付与された管理記憶識別番号7000番を有し、管理記憶装置15の物理記憶領域毎に7000個の管理記憶識別番号が割り当てられている。

にあるかを認識することができる。

【0067】図10は物理記憶装置15内の稼働情報取得部223の占有時間集計部227に累積・格納される1/10による記憶領域へのアクセスによる占有時間情報のテーブルの構成例を示す図である。

【0068】このテーブルは、編組配位情報番号801
毎、1/〇 JOB種別802番に占有時間を集計した
ものである。1/〇 JOB種別として、図示例では、
シーケンシャル1/〇8010、シーケンシャルデータ
ータ820、シーケンシャルラバリティ830、ラ
ンダム1/〇840、ランダムラバリティ860、
キャッシュヒット870、合計880が示されている
が、さらに他の1/〇JOB種別があってもよい。

【0069】物理記憶装置15内のデータ処理制御部52は、ホストからの1/0等により記憶装置10にアクセスすると、アクセスがあった物理記憶領域801に
関して、アクセスの108識別802年に、これより、900を占有時間累計部227に累積する。これにより、物理記憶装置15内の物理記憶領域の番号と、その物理記憶領域へのアクセス回数による占有時間と、ある時間内の合計占有時間との関係を物理記憶装置15内で得ることが可能となる。

【0070】前述した本発明の実施形態によれば、1／0毎の記憶装置の占有時間情報を取得することができ、かつ、記憶領域へのアクセスによる記憶装置の占有時間情報の取得を物理記憶装置内で実現することができ、

【0071】本発明の実施形態によれば、前述により、ストレージサブシステムを構成する情報の物理記憶領域と管理領域の占有時間を取得することができ、物理記憶領域の管理記憶領域の占有時間のみでは取得不可能なシステム構成で各論理記憶領域の占有時間（実稼働時間）取得することが可能となる。また、前述した本発明の実施形態によれば、1つの1/0毎の物理記憶装置への影響を軽減することによって、前記ストレージサブシステムの論理記憶領域の利用率の節約や利用効率の向上を小さくすることができる。より高度な性能チューニングを行うために、より精度の高い物理記憶装置への1/0毎のアクセス占有時間情報の取得を行うことができる。

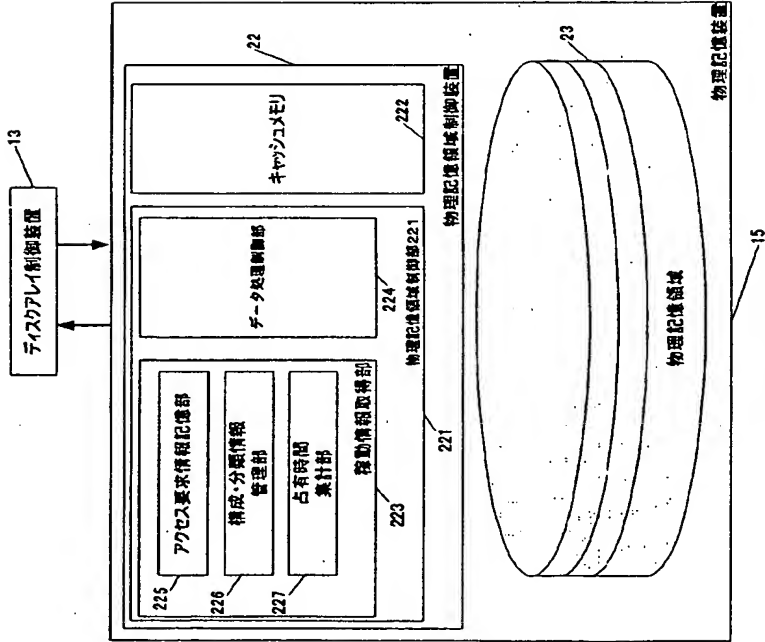
【0072】

【発明の効果】以上説明したように本発明によれば、物理記憶装置が論理記憶領域の占有時間を取得することができ、ディスクアレイ制御装置のみでは取得不可能なシステム構成で各論理記憶領域の占有時間（実稼働時間）を取得することができ、

【0073】また、本発明によれば、物理記憶装置への1/0毎のアクセス占有時間情報を取得することのできるため、1つの1/0毎の物理記憶装置への影響を考慮したストレージサブシステムの論理記憶領域の利用率的な

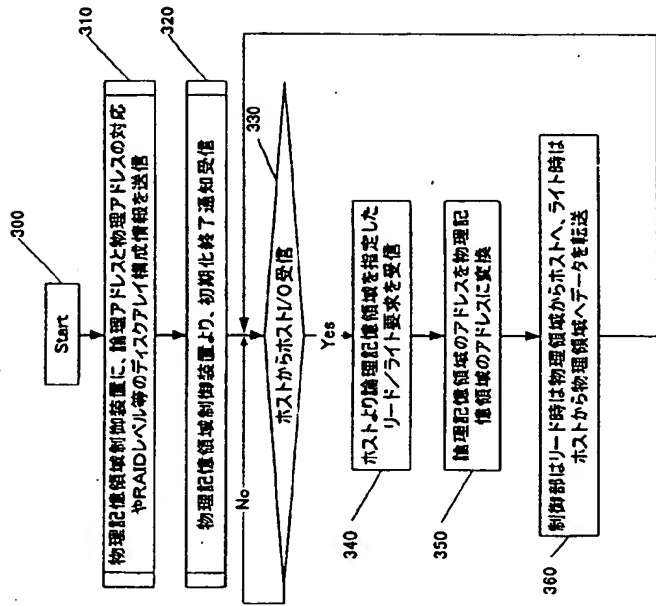
【図2】

図2



【図3】

図3



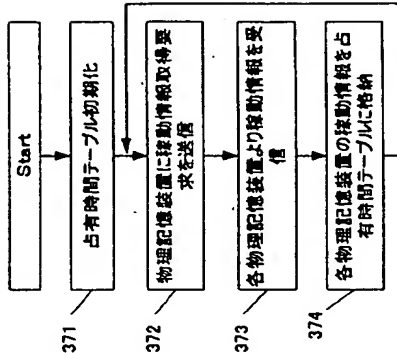
【図9】

図9

物理記憶領域番号	物理アドレス	物理記憶領域番号	物理記憶領域アドレス
1	3000~3999	1	0~999
4	4000~4999	1	1000~1999
5	5000~5999	1	2000~2999

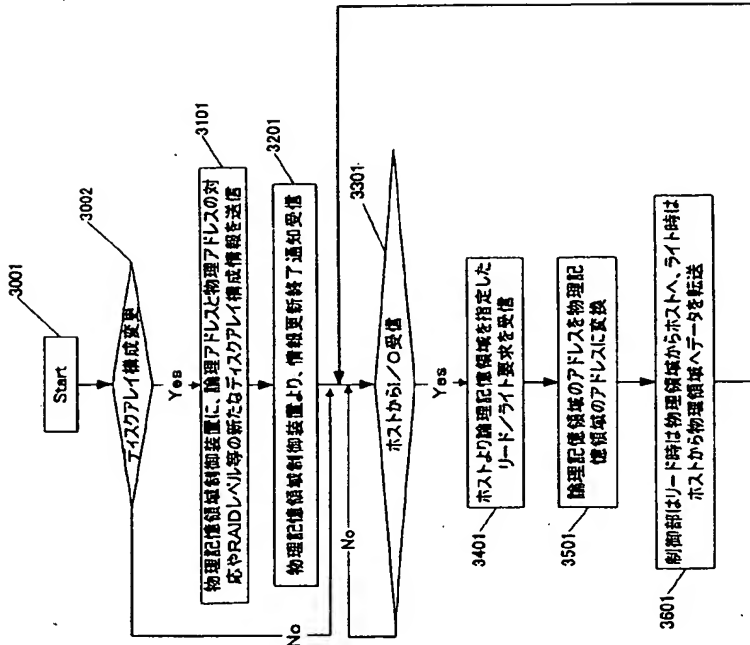
【図5】

図 5



【図4】

図 4



【図7】

図 7

物理記憶装置番号	物理アドレス	物理記憶装置番号	物理記憶領域アドレス	レイドレベル	パリティグループ番号
0	0~999	0	0~499	1	100
1	1000~1999	0	1000~1999	1	100
2	2000~2999	0	2000~2999	1	100
3	3000~3999	1	0~499	5	120
4	4000~4999	1	1000~1999	5	120
5	5000~5999	1	2000~2999	5	120
...

